September, 1991

A PROJECT FOR THE STUDY OF HUMAN DIVERSITY WITH SPECIAL ATTENTION TO VANISHING HUMAN POPULATIONS

HUGO Committee for Human Genetic Diversity

Chairman: L. L. Cavalli-Sforza, Genetics, Stanford, CA, USA
Members: J. Bodmer, Genetics, ICRF, London, England
K. Kidd, Genetics, Yale, New Haven, CN, USA
M.C. King, Epidemiology, Berkeley, CA, USA
S. Paabo, Genetics, Munich, Germany
A. Piazza, Genetica Umana, Turin, Italy
M. Siniscalco, Genetics, ICRF, London, England

Table of Contents

A. INTRODUCTION AND SUMMARY

This project has two important, complementary aims.

1) preservation of crucial biological information on vanishing human populations and on a selection of current major population groups and

2) analysis of human diversity in conjunction with the current human genome project.

Molecular methods which are now being applied to the study of the human genome are revealing a level of diversity between individuals which is far greater than had been detected using previously available techniques.

As a result, the precision with which populations and their origins and interrelationships can be defined using relatively small samples has increased enormously. The population samples that we propose are collected and preserved will

a) be an invaluable cultural resource for establishing the history of modern population groups and their precursors;

b) provide base line information for the investigation of genetically based variation in disease incidence between populations and for identification of individual samples for paternity, forensic and other applications;

c) enable studies on the functional basis of polymorphisms for expressed genes specially with respect to the actions of natural selection.

This is a project which requires international collaboration and which will provide fundamental information that will be freely available throughout the world to any interested researcher. It is therefore entirely appropriate that this project should be promoted by HUGO alongside its other international activities.

It is difficult - if not completely impossible - to study individual variation for the whole genome. The following selective strategy is proposed.

1. For rare and vanishing populations collect data on variation in a suitable sample of neutral genes to define the population and its contribution to other populations. 'Neutral' variation is most easily defined operationally as that occurring basically in pseudogenes, introns, flanking regions and third base pair positions.

2. Using these data as a background study functional genes in the same populations looking for differences which could be relevant to the present population structure.

3. In addition to the rare populations it is essential to study the major large current populations since it is these which are relevant to future medical and economic planning.

Unfortunately just as we are acquiring the means of analyzing the whole human genome efficiently, the aboriginal populations showing the greatest diversity, whose study is most important for understanding human history, are disappearing.

The importance of specific populations to problems of human evolution is enormous.Examples of four such populations are described later.

A summary of major research steps

1) Select a set of aboriginal populations representing the most interesting ones from an evolutionary point of view, choosing especially those more likely to disappear because of loss of identity or other reasons. They will be sampled for the purpose of growing and storing permanent cell lines from a certain number of individuals from each population, in order to provide eventually an unlimited source of DNA.

2) Organize and support the collection of blood samples and their shipping to laboratories and banks for storage and transformation. Two or three major cell banks, say in the United States, Europe and Japan should store aliquots of each of the cell lines.

3) Set up laboratories in different parts of the world, supporting existing ones where possible, for the transformation of cells, and the preparation of DNA. Population studies, allied to specific diseases of local interest, can form a valuable component of Human Genome projects for less well developed countries. Through this they can establish contact with resource centres providing cell lines, DNA probes and other materials and also with appropriate international data bases. Such studies can also form a useful basis for training in techniques, data analysis and data base handling, which are the essential requirements for the medical, agricultural and other applications of Human Genome analysis.

4) Ensure that all DNA samples, viable cells if required and data are available to research workers interested in testing or in analyzing results.

5) Test batteries of markers and regions representative of the genome, carefully selected for efficiency and reliability, as well as for functional interest (e.g.HLA).

6) Collect material other than blood e.g. cells from a mouth swab or a hair follicle from some populations where blood sampling may be difficult or impossible. Such material may eventually be expanded efficiently by, for example, PCR - based micro-cloning techniques.

7) Set up research, for example, on techniques of white blood cell separation and conservation, T lymphocyte and other cell immortalisation methods, generation of new and economic markers, collecting strategies, testing techniques, and efficient clone library production using limited amounts of material.

8) New techniques for studying at least some DNA markers in ancient material, including especially bones, open up the possibility of working with archaeological material. This could be of great interest for comparisons with existing populations (e.g. pre Columbian and present day American Indians)

suggesting the possibility of setting up, in addition, a bank of such material for DNA studies by arrangement with Natural History Museums and other sources of archaeological palaeontological collections.

9) As an important addition to the collection of vanishing populations a strategy should be worked out for the collection of samples from major ethnic groups around the world.

10) Develop a data base for the population material and derived genetic information. This should be integrated with existing data bases such as CEPH, OMIM and GDB.

A small meeting of the CIBA type should be arranged as soon as possible to include palaeontologists, anthropologists, linguists, geneticists and molecular biologists to discuss the project in general and this document in particular.

The discussions at such a meeting should help clarify the project and set priorities for the populations to be sampled, technical approaches and research strategies. A minimum of 200 and a maximum of 500 populations should be selected, with wide geographic spread and aimed at maximising the range of populations sampled with respect to their presumed ethnic backgrounds. A sample of 50 random individuals from most populations would be selected, though larger samples could be drawn from representatives of the current major world population groups. Blood banks and tissue typing centres could be very helpful in this respect. It would also be important to seek collaboration from the laboratories doing population studies as part of the current 11th International Major Histocompatibility Testing Workshop.

Collection and Testing Strategies

a) Collection of blood samples

An anthropologist and a doctor, preferably with local knowledge should form core of the team collecting samples. Three types of sample may at first be collected.These would be for immortalisation using EBV transformation, for long term storage and for immediate preparation of DNA.

b) Test of selected batteries of markers

Testing selected batteries of markers is essential. The best analysis is done on a set of data in which every population has been tested for the same markers. However some selection of neutral markers that show large frequency differences only between subsets of populations may be appropriate.

A great advantage of the essentially unlimited range of polymorphic DNA markers that can now be studied is that their sheer number increases the chance that one or more 'neutral' polymorphisms will differ substantially in frequency between populations of different origins. The larger the frequency difference, or variation, the more informative is the marker for a given sample size.

c) Blood-sample collections already in existence, or new collections of type C only (without immortalization)

There already exist some important collections of blood samples (serum, plasma, etc.) that could be made available for this project. There are two main limitations to this: their existence is known to very few people; retrieving the samples, selecting, aliquoting and shipping them to interested laboratories may be expensive for the laboratories when they are stored. Both limitations could be probably removed with modest financing.

Budget

The budget will include money for expeditions, immortalization, and storage, and initial testing of a core of essential sequences and polymorphisms for the first five years. The logistics of the distribution of cells and DNA to laboratories, and the generation of a data base for collecting all results and making them available on a near-real time basis, will have to be discussed in appropriate international meetings.

B. GENERAL BACKGROUND

Molecular methods which are now being applied to the study of the human genome are revealing a level of diversity between individuals which is far greater than had been detected using previously available techniques. The probability that a given nucleotide will differ between two random individuals is roughly estimated as between 1 in a 100 and 1 in a 1000, though this varies considerably depending on the function of the DNA segment considered;

Changes in functionally important regions are likely to have pathological consequences. Segments involved for example in basic cell metabolism (housekeeping genes) therefore show less variation between individuals whereas certain genes like HLA may have a relatively high level of variation. The causes for variation are not completely understood but may be related to protection against pathogenic agents. Thus, the search for individual differences is potentially useful for physiological and medical reasons.

It is inevitably difficult - if not completely impossible - to study individual variation for the whole genome.

Analysis of individual variation of the genome at the DNA sequence level is so far extremely limited. Thus sequences from more than one individual exist for only a few DNA segments. Data for the study of restriction fragment length polymorphisms (RFLPs) have generally been collected from poor samples, usually a small number of individuals who are readily available to research workers and who are mostly, but not necessarily, all Caucasoid. These samples are adequate for an approximate prediction of the usefulness of a given RFLP for linkage studies in Caucasoids, but they are of very little use - and in some respects even misleading, for an evolutionary study of humans as a whole. Analysis of variation should be carried out on a well-chosen set of individuals, suitably representing human diversity and matching the particular purpose of the investigation.

Unfortunately just as we are acquiring the means of analyzing the whole human genome efficiently, the aboriginal populations showing the greatest

diversity, whose study is most important for understanding human history, are disappearing. The numbers of individuals in most ethnic groups of the greatest interest for human evolution are decreasing rapidly. Many individuals of such groups are losing their identity because of migration to cities or, more generally, areas where jobs are available; because of increased admixture with old or new neighbours, high mortality, and, not infrequently, high sterility, usually of unknown cause. We may have no more than a decade or two before many of these groups disappear.

The populations most interesting for studies of human history and evolution general show the highest genetic diversity. They live in the most remote places and are most at risk of vanishing, as many already have, for example, the natives of the Caribbean, Tasmania, and Tierra del Fuego. All are dead in the first two cases, and only a few hybridized individuals survive in the last. Many populations can be identified as potentially interesting because they have a unique appearance, have lifestyles typical of the past, speak languages that are extremely rare and nearly extinct and that may even be unlike any other known languages. There is an obvious correlation between geographic remoteness, physical or cultural uniqueness of populations, their interest from the point of view of human history, and their danger of vanishing for one reason or another.

The importance of specific populations to problems of human evolution is enormous. To cite a few examples: In the Andaman islands live four groups of Negritos, three of which are now reduced to about 100 individuals each, and the fourth to 20 or so, who were flourishing until 50 to 100 years ago. Disease, alcoholism, and opium (spread deliberately in earlier times) have practically destroyed them, and they have almost no children. Some of the commoner reasons for infertility, for example venereal diseases, and diet, have apparently been discarded, and it is possible that sterility is due to excessive inbreeding within each group. These populations could be the last remnants of those that migrated from Africa to Australia. In India, and in mainland and insular Southeast Asia, there are tribal populations that, although different from the Andamanese and other Negritos, might be similar candidates. In a tropical forest

west of Addis Adaba - a very unusual environment in Ethiopia - there are groups that may have some genetic similarity to Khoisanids (South African Bushmen), indicating an ancient presence of these people in Ethiopia (this has also been suggested by some research workers on other grounds). In many mountainous areas are populations who have been there a very long time and may represent some of the most ancient aborigines. The Caucasus is one such area; languages spoken there are claimed to resemble Basque (Spain), Sino-Tibetan languages, and Na-Dene languages spoken in British Columbia. If this is true, they probably belong to a very old linguistic superfamily that has been widely replaced almost everywhere by later expansions of people speaking other languages. Burushaski, a language with no well-ascertained affiliation, is spoken by the Hunza tribe in the mountains of northern Pakistan and may indicate that its speakers are survivors of an ancient group, like Basques. There are a few other language isolates like these, especially in Asia. This is only a very small sample of mysteries to which genetics can bring some light.

C. SOME SPECIFIC RESEARCH ISSUES

1) For populations for which it is difficult to collect enough blood samples for transformation, and for chromosome regions that are destroyed in transformation of B lymphocytes, DNA can be obtained in limited amounts directly from blood or other sources and stored for future work. Also to be considered here are other collections of biological samples that are already in existence or could be made independently. With modern techniques, it is not necessary to use blood: for example serum, plasma, hair roots and mouth swabs can be used for certain purposes. Many other sources of small amounts of DNA can be useful. Libraries can be prepared from small amounts of material and increasing the efficiency of doing this should be an aim of the research. A computer catalogue, and/or a repository of already existing and new samples should be made available to research workers.

2) As an important addition to the collection of vanishing populations a strategy should be worked out for the collection of samples from major ethnic groups around the world. Collaboration with blood banks, tissue typing or forensic laboratories to collect random specimens should be developed. Where possible these samples should be taken from regions where they have lived for some time.i.e. European centres for European groups rather than from regions to which they have recently moved, e.g. the U.S.A., to reduce the problem of possible admixture.

D. SOME DETAILS OF RESEARCH STEPS

Meetings should be arranged as soon as possible to focus on the modalities and priorities for collecting samples, feasibility of the proposed studies, and planning a programme for the technical developments which should be supported.

1) Choice of populations

The number of populations to be sampled and the sample size will be decided on the basis of the opinion of specialists in population genetics, in a meeting soon to be convened. A possible strategy is the following: the number of populations should be not less than 200, and possibly as large as 400 or 500. Forty to fifty individuals may be immortalized per population; samples from more individuals will be added, to be stored without immortalization (ideally in a condition that allows later immortalization).

The choice of populations should follow certain rules, to be agreed upon. Possible rules might include the following:

a) In the case of the vanishing populations, people and regions chosen for sampling should give adequate guarantees of representing reasonably aboriginal populations. Cities should be avoided, unless groups collected there are made of very recent immigrants whose origin is well known and clearly desirable. Most interesting areas will be sufficiently isolated that the chance of miscegenation is reduced.

b) The remote origin of the people may be unknown, but there must be a good reason for specific interest in them. A good history of each individual should be collected, but the identity of the individuals sampled will remain confidential. If some interesting finding suggests that it might be useful, the possibility of resampling the same individual(s) could be considered with appropriate precautions.

c) For each population, individuals will be chosen sufficiently far apart in degree of known relationship and geographic distance that the chance of close relationship is small. It may be difficult in general to follow a specific grid. In mountainous areas, which are among the most desirable, a natural criterion is to sample different valleys equally. The criteria for choosing populations, and within them individuals of adequate diversity, should be tempered by the following constraints (d, e).

d) For vanishing populations that are very small but of great importance - e.g. Andaman islanders - all living individuals should be sampled.

e) Populations that are especially critical should be given priority, but even if a sample of populations smaller than 500 were to be studied, they should represent the whole world as homogeneously as possible. Thus, one should start with widely distant populations and fill the gaps gradually.

A list of world populations that includes in considerable detail most of those of interest for our purpose, and is reasonably up-to-date with respect to numbers of individuals and locations can be found in The Ethnologue (editor: Grimes). It will be necessary to rely as much as possible on information from anthropologists who have recently visited the group(s) to be studied.

Vanishing populations will certainly be given priority, but the world sample should also include populations that are reasonably "aboriginal" and yet are not, strictly speaking, vanishing, so that the sample will represent the world. These will include the major ethnic groups, Europeans, Han Chinese, Japanese, Indian and African groups. Populations in the Asian Soviet Union (or Federation) will also be studied. The usual rules of confidentiality will apply.

2) Collection of blood samples

An anthropologist and a doctor, preferably with local knowledge should form the core of the team collecting samples. Three types of sample may be collected.

a. Cells for immortalisation. In general a total of about 50 male and female samples will be taken from any population except those very small groups where all individuals will be collected for immortalisation.These should be collected on the last day of the visit if possible or in any case transported to the laboratory as quickly as possible.For the 200-500 populations envisaged this will mean a total of 10,000 to 25,000 samples to be immortalised over a period of 5 years.

b. Samples not to be immortalised immediately. These should be separated and stored at low temperature in liquid N2 to retain viability for possible future immortalisation if required.The logistics of field storage and transportation will need to be worked out.

c. Samples for extraction of DNA only. These samples can be kept at +4°C for a few weeks until required or transported.Other sources of DNA may also be investigated.

As already noted, there already exist several collections of blood and serum samples, and their potential use for our purposes is discussed below in point 6.

In general, the most interesting populations may be sampled in all three ways.

A crucial question is whether families or individuals should be sampled. For establishing evolutionary history, families are redundant, and one would try to select individuals as unrelated as possible, as indicated above, but always within a given ethnic group. This will maximise the pool of genetic material available from an ethnic group. On a practical note, since family relationships are not always clear in some populations and also because some samples may be lost,

each individual sampled will be independent of his kinship with another individual and of the necessity of the other sample surviving.

3) Transformation of B lymphocytes

Immortalizations have a high proportion of success provided that blood samples reach laboratories within 24-48 hours of collection. The logistics of organising this and the possibility of extending that time span could usefully form part of the technical research to be undertaken.

Because of these limitations and in any case to stimulate and maintain collaboration throughout the world, it is necessary to establish or support existing laboratories in critical areas of the world for immortalising local populations. Laboratories with demonstrated interest and competence already exist in many parts of the world including several centres in Europe, Japan, South Africa, China, Australia, Russia, India and the United States.

Many important aboriginal populations are still living in China, India, and the USSR. Their total number of inhabitants comprise nearly half the global population. All three are very rich in extremely interesting aboriginal populations. It should be easy to establish a small laboratory for immortalization in each of the following places:

a) Beijing, possibly at Academica Sinica, where LCS has collaborated with Prof. Du Ruofu, who is already carrying out a genetic survey of the 55 ethnic minorities of China for non-DNA markers;

b) Delhi, at ICGBE, a new molecular biology laboratory organized by UNIDO, a branch of the United Nations, where good-level molecular biology has already been going on for a couple of years;

c) Moscow, and/or also Novosibirsk, where there are several possible choices of laboratories that already have a long tradition of population studies, but probably not of immortalization.

This leaves out areas in Southeast Asia and the Pacific Ocean, where extremely interesting populations also exist. There would be no shortage of good laboratories in Australia. A centre that could serve mainland and insular Southeast Asia (Indonesia, Malaysia, Borneo, the Philippines, etc.) could probably be located in Singapore. Other possibilities also clearly exist, e.g., in Pakistan (to serve also Iran, Afghanistan, etc.)

In the United States, two laboratories (LCS and Kidd) are already involved in growing out cell lines. The Coriell Institute (Camden, N.J.) which is prepared to deal with a large number of new cultures and is flexible in its organization has been approached for help (e.g., in Capetown and Johannesburg).

Europe is, of course, the least difficult to organize, and because of the good network of air connections, European laboratories could also take responsibility for Africa and the Middle East. There are some possibilities for doing part of the work in Africa.

The new laboratory being developed at Alghero, Italy, to be directed by Marcello Siniscalco might prove to be a valuable central resource in Europe. The matter of transport connections however would need to be clarified.

Most or all of these laboratories should also do amplification of cell cultures for DNA extraction, which should be sent to the central laboratories for quality control and distribution. The question of whether it is convenient to extract DNA locally depends on the quality of results. Some smaller centers collecting blood samples and transforming lymphocytes may prefer to ship cultures immediately after transformation, instead of sending DNA. In principle, all cell cultures should be maintained in liquid nitrogen both at the center of production and at two or more repositories, in order to minimize chances of loss

and to facilitate shipment around the world, particularly where import of cultures into the U.S.A. can cause problems..

Storing a copy of every cell culture in a United States repository demands that cultures be grown in fetal-calf serum that has been accepted by the Department of Agriculture and, in practice, produced in the United States, in order to avoid the spread of cattle diseases. Artificial substitutes for fetal-calf serum have worked well for us and are much cheaper but may be less efficient. General use of DA-certified fetal-calf serum would be best, even if it is far from cheapest. Problems of this kind will require attention.

A good United States laboratory with one technician and a part-time supervisor could do a maximum of 400 immortalizations a year at a cost for reagents of $100 per individual including amplifications, plus the salaries of the technician and the part-time supervisor (figures from Judy Kidd). This would correspond to the immortalization of 8 populations a year, 50 individuals per population in a (well-organized and active) single laboratory.

4) Distribution of DNA to research workers

The choice of a great variety of markers will also be assured by making DNA available to interested research workers either as part of an accepted collaboration free or for a fee to cover costs.

The Coriell Institute at Camden, N.J., is prepared to handle the requests for DNA and cultures, starting from cells, blood, or perhaps other types of samples. As mentioned before, it already has a small list of "human diversity cultures" that has been provided from the joint Stanford-Yale collection; only a few individuals per group are available and will appear for the first time in their September catalog. The current charge is $50 for micrograms of DNA per individual. For PCR, much smaller samples of DNA (1 microgram or less) should

be distributed, and the charge should be much less. The cooperation of other centres in Europe and Japan should be sought.

If samples are to be made available without charge or at low cost, it will be necessary to establish rules for accepting participants in the programme and for determining the minimum number of individuals and populations to be examined. The example of CEPH is a very useful precedent.

In any case distribution to any laboratory will be dependent on a contract stating that the recipient laboratory will make available, in standard format, the results of the tests carried out on the samples. These data will be added to the database.

5) Test of selected batteries of markers

Testing selected batteries of markers is essential. The best analysis is done on a set of data in which every population has been tested for the same markers. Data on classical (non-DNA) polymorphisms examined on human populations are extremely sparse; only a very small fraction of all the data available is really useful, because (except for HLA) there never was a systematic testing of a selected batch of populations, done by a number of scientists using the same markers and techniques. If the choice of markers and populations were left entirely to the research workers, the results would be extremely sparse and would not allow a reasonable set of results to accumulate in a reasonable time. This is not to say that the material should not be made available to any laboratory who wishes for its own reasons to test 'non standard ' markers. This might however be charged at a higher rate than to those laboratories performing the standard 'service' testing.

The best strategy would be to:

(1) ensure, by some type of contract, that in some laboratories a systematic testing of certain markers of importance is done on all populations, and

(2) ensure that DNA is distributed free or at reasonably low cost to laboratories that volunteer to do certain analyses of their choice. When published, all results should enter a data bank available to research workers at most at a fee to cover basic costs. This data bank will form part of the set of data bases associated with the Human Genome project.

The number and type of markers to be tested on all data may be decided while the collection of samples is proceeding. Part of the data will be sequences of chosen segments; others will be polymorphisms of various kinds. Experience in a pilot project using RFLPs is that 100 polymorphisms is an absolute minimum and depending on the level of polymorphism of the markers chosen may not be sufficient for all types of problems; 200-400 may be necessary for testing an extended range of hypotheses.

The analysis of sequences may be carried out in many different ways. A minimum could be the analysis of one individual per population, but it is hoped that with the development of more efficient automated sequencers, the number of individuals per population could be increased. The information gained by increasing the number of different segments to be sequenced on only one individual, or by increasing the number of individuals per population but using a smaller number of sequenced segments, needs to be established.

The advantages over RFLPs, and the problems presented by mini- and micro-satellites, etc. are discussed in the paper by Bowcock and Cavalli-Sforza (1991, Genomics).

Much preliminary work must be done in order to choose an adequate number of sequences to be studied. Today, there are a few sequences in mitochondrial DNA and in HLA that have enough individual variation that it is worth studying them systematically. Many more should be chosen on the basis of criteria still to be identified.

It is too early to say what level of testing efficiency will be possible 2 or 3 years from now. The testing of polymorphisms and establishment of sequences is steadily being automated; before starting systematic testing on a large scale, it may be more efficient to wait until the technology that is now being developed for the Genome project is closer to becoming operational. For the present, there are five possible areas of activity.

a) We can continue a reasonable level of testing with batteries of RFLPs that have already been examined on a certain number of populations, so that the addition of information on new ethnic groups can be added to existing knowledge.

b) We can continue sequencing short regions, such as the control region of mtDNA and some other sequences. This could also be very useful as a stimulus for theoretical work to improve methods of comparisons of sequences, which are now of doubtful efficiency from an evolutionary point of view.

c) We can organize good batteries of polymorphisms with the newer techniques such as PCR (polymerase chain reaction), SSCP (single strand conformational polymorphism), and DGGE (denaturing gel gradients), that increase the efficiency of the analysis.

d) We can continue the search for sequences that are useful additions to the few already known and can be used for a systematic testing of variation of a sample of DNA segments that can be considered representative of the genome and have useful prerequisites for evolutionary analysis.

e) We can expand the range of HLA typing by oligonucleotide probes which is already started for the 11th International Histocompatibility Workshop held in Japan in November 1991. This is a good forum for stimulating interest in this programme since many laboratories represented there have a great deal of experience in the collecting, shipping, preserving, transforming and testing of large sets of samples and an already demonstrated interest in population diversity.

6) Blood-sample and DNA collections already in existence

September, 1991

There are important collections of blood samples (serum, plasma, etc.) that could be used if of interest. There are two main limitations to this: their existence is known to very few people; retrieving the samples, selecting, aliquoting and shipping them to interested laboratories may be expensive for the laboratories where they are stored. Both limitations could be removed with modest financing. Examples of such collections are: the samples of white cells gathered by HLA researchers; the Yanomama and other southern Amerind populations collected by J. Neel and his coworkers and entrusted to Ken Weiss of the Pennsylvania State University; the collection of over 1000 African Pygmy samples by LCS, currently at Leyden University in the laboratory of L. Bernini.

A possible solution should include the following phases:

a) Through appropriate appeals in scientific journals, obtain the information on available collections: sources, number and average size of samples, and their type (blood, plasma, serum etc.); data available per individual; intention to donate the collection, or ability to effect distribution of samples to interested parties. After obtaining estimates of cost, one could concentrate collections that the owners are willing to release at Coriell, and other comparable centres in Europe and Asia.

b) For the collections thus donated, appropriate aliquots of samples - or in some cases DNA extracted from them - could be sent by Coriell or by owners to interested research workers.

c) These samples may have a short life, if they are unexpandable or irreplaceable by current techniques. A permanent committee could examine (by mail) requests in order to decide which should be accepted and which have priority. This decision would, of course, be necessary only for samples stored at Coriell and other major centres; one may also want to give a vote (or a veto power) to the original owner of the samples.

d) Make information on the collections, the quantities of samples, and the conditions for their release available to interested research workers by appropriate bulletins on computer networks.

7) Research needed (see technical appendix and introduction)

1. Improved methodology for systematic testing for diversity

2. Improvements in short term storage and shipping methods for samples.

3. PCR libraries should be generated, particularly in cases where source material is scarce or limited

4. Theoretical investigations of optimal sampling strategy should be undertaken.

E. DEVELOPING A BUDGET ESTIMATE

1) Meetings on how to sample, whom to sample, and the general organization of the program should be convened.

2) The cost of collecting samples and transporting them to the laboratories can vary widely. In countries that have a reasonable system of internal air services and the possibility of renting cars or in which help from the military can be obtained, the cost of collecting samples is relatively small. This has been the case in USSR as reported to LCS by Russian research workers with whom he has been in contact, and LCS hopes to obtain support of the Chilean navy in Tierra del Fuego. In very special circumstances, in which helicopters or other special means of transportation must be hired, the cost would be very high, but this should be applied only in very special circumstances. Substantial work could be done by anthropologists who are already engaged in studies of populations for their own interests. All their extra costs should be paid. This might guarantee that the sampling would be done by scientists who are well acquainted with the populations in question, and reduce costs.

The estimate should be divided into two parts:

a) For laboratories that do their own immortalization, the cost should be included in their expenses (see next point 3). The expenses are for domestic trips, which are ordinarily cheaper in the countries in which such laboratories operate (e.g., China, India, Russia); it may be easy to obtain transportation from the local army or air force at zero cost. This was true of Russia in the old regime.

b) For anthropologists or geneticists who apply with specific programs and send their samples to Coriell, the cost should be considered here. The cost of one expedition may vary from a minimum sum of say $2000 to much larger sums. We calculate an average cost of $6000. A doctor and an anthropologist may

be a minimum unit. Medical help may also be obtained locally and, in several Third World countries, local anthropologists are certainly the best choice.

3) Much immortalization work would be done at Coriell or other major centres, but only from locations that are easily accessible. Especially for East, Southeast, and South Asia and for Oceania, much work could be done more profitably in local laboratories, unless there is a radical change in present techniques of immortalization. The cost of immortalization and DNA extraction at Coriell for this new endeavour is not known, and needs to be established.

The cost of starting one local laboratory dedicated to immortalization, cell amplification, and perhaps DNA extraction of samples obtained in the country in question is small. The cost of equipment is around $30,000 (including a centrifuge, CO_2 Incubator, laminar-flow hood, spectrophotometer, roller-bottle apparatus). Depending on the country and local rates of compensation for personnel, the yearly costs may be less than $100,000 per laboratory. This should include, at least in some cases, travelling and shipping expenses for collecting samples (unless they demand special travel conditions), salaries, reagents, and communication expenses. This calculation would set a minimum cost at $200 per immortalized individual.

Ideally one would like to collect 200-500 populations. Clearly, collection is the primary phase and should be the focus of the first years. If a minimum of 10,000 individuals (40 per population, 250 populations) is to be immortalized, the minimum cost of immortalizations would be about 6 million dollars, to be distributed over 5 years. This, however, is a minimum, which might well have to be increased considerably, and perhaps doubled.

Distribution of work from the various continents is planned on the basis of minimizing the number of necessary air connections, which are the most frequent cause of delay and consequent loss of cultures. Roughly estimated, the work might be distributed as follows (number of populations based on a maximum of 500):

a) work in the Americas (120 populations) to be done partly at Coriell and partly at Yale and Stanford, and perhaps a South American laboratory to be considered;

b) work in Europe and West Asia (50 populations) at Alghero, Sardinia (where it is hoped that a laboratory directed by M. Siniscalco and supported by the Italian government will start functioning in 1992);

c) work in Africa (80 populations) to be done in part in Africa but mostly in Paris, or possibly Amsterdam, or London, which are reached by direct flights from Africa and Western Asia;

d) work in Asia (150 populations) to be done in the USSR for Northeastern Asia (and much of European Russia), in China for China, in India for India, perhaps in Singapore for Southeast Asia; also in Japan.

e) work in Oceania (100 populations) to be done in Australia New Guinea, and/or Singapore.

Perhaps 7-10 local laboratories might be set up, 4-5 in Asia, 2 in Oceania, 1-2 in Europe,one in South Africa and 1-2 in South America.

4) If research workers pay costs for DNA sent by the repositories, including DNA provided from local laboratories around the world, then there will be no cost for this item. If it is provided free of charge or at reduced cost, then an item should be entered in the budget reflecting this cost. It is difficult to evaluate, and therefore has not been included at this stage.

5) It is essential to organize systematic work on markers, but it may be premature to plan now any major extension of the work already done with DNA markers on immortalized populations. As already mentioned, the relevant technology is undergoing a profound transformation under the stimulus of the genome projects, so that starting a large-scale screening of populations now may

be premature. In any case, the number of available populations is currently very low, and it will take 2 or 3 years before a substantial number is accumulated. It seems reasonable to delay this part of the program by 2 or, at most, 3 years, dedicating more effort at the beginning to the primary objective namely, to the collection of populations, and the search for techniques and segments to be tested systematically. Ongoing investigations, testing batteries of already established markers on populations being collected could be profitably continued. New batteries of markers on a limited number of populations could be started. They will be useful as pilot studies for future, more ambitious research, using a larger number of populations. The testing of a large number of populations will be possible only when more collection work has been completed.

6) The cost of storing and distributing DNA from blood samples collected earlier, or collateral to the collection of blood samples for immortalization, should be relatively small, but is difficult to evaluate. The Director of the Physical Anthropology Division of the National Science Foundation, Dr. Mark Weiss, has indicated a special interest in this program, which should not, however, exclude other modes of participation by NSF.

References

Bowcock, A. and L.L. Cavalli-Sforza (1991), The study of variation in the human genome. Genomics, 11:491-498.

Grimes, B. (19??) The Ethnologue. Wycliffe Bible Translators, Inc. Dallas, TX.

F. TECHNICAL APPENDIX

1. Pilot study

New methodology developed for the study of genetic diseases at the molecular level has radically changed the possibility of conserving the DNA of these populations for tests that are hundreds of times more powerful than those available until now. This approach has already been applied in preliminary investigations of aboriginal populations by L. Cavalli-Sforza of Stanford, in collaboration with Ken and Judy Kidd of Yale. Fifteen populations have been sampled, the first two of them personally by LCS, others by collaborators, and an approximate average of 50 individuals per population has been immortalized in the two laboratories (5 populations immortalized at Stanford and 10 at Yale). A list of these is given in an Appendix. In the original plan of the Stanford-Yale project, a small sample of these cultures (also listed in the Appendix) have become part of the catalogue of the Coriell Institute (Camden, N.J.) to be published this month, and will be available to scientists. Cells of each individual have been multiplied to produce 500-1000 micrograms of DNA, and five populations have been examined for at least 100 RFLPs (restriction fragment length polymorphisms), 83 at Stanford and 17-30 at Yale. Another five populations have been tested at Stanford for 40 of the standard 85 markers, and the work is being extended to the other markers. This research can be considered a pilot project for the present program, but was designed on the basis of molecular techniques available in 1985. There has been an enormous increase in the power of the molecular approach because of new methodologies, like the polymerase chain reaction (PCR) and the automation of sequencing. Clearly, today one must think of a thoroughly new approach that takes into account the rapid evolution of the methodology but assures the storing of the human material before it is too late.

2. White and other cell preservation

There is need for more research on techniques of white blood cell conservation, generation of new and economic markers, collecting strategies, and testing techniques. New approaches to immortalizing T lymphocytes, fibroblasts and other cells should be investigated.

3. Transportation and extension of viability of samples

Cells for transformation must be kept at room temperature in transit. This generates serious logistic problems in certain areas. One can say that most collections of samples in America have been successful when sent to North American laboratories, and the same is true of African samples sent to European laboratories. When sent to North America, African samples enjoyed much more modest success because of the longer time of transportation. Transit time is critical. There have been claims that freezing whole blood may be enough for transporting cells safely to the laboratory (Chenevix-Trench et al. 1990, AJHG 46:635), but attempts to repeat the technique have failed thus far (Dr. R. Mullivor, pers. comm.).

4. Collection of blood samples

To organize the collection of blood samples from a specific population, ideally a minimal team of an anthropologist and a doctor should spend sufficient time on site; collection should, however, be made on the last day, if possible, so that samples can be transported rapidly to the laboratory. Before the final day, one may collect samples for conservation without immortalization, as well as data regarding individuals and populations. In some cases, lymphocytes may be stored in liquid nitrogen on site and transported to the lab for immortalization.

The collection of samples should obviously be authorized by the authorities of the country in question and the anthropologist should ideally be familiar with the population, or at least related ones. Many anthropologists know how to collect blood samples, and a doctor would not strictly be necessary, but medical

assistance within the limits of means available and an analysis, however superficial, of the local state of health would be important and necessitates a doctor. In some circumstances, it might be better if the help of a local doctor were secured, as it is often important that he/she be familiar with the local pathology.

5. Other sources of DNA

Buccal scrapings, hair roots, serum and plasma are all possible sources of DNA. The use of discarded material such as placentae should be investigated.

A possible alternative is to carry into the field equipment for washing white cells and storing them in liquid nitrogen; but air shipment of tanks of liquid nitrogen is difficult. Airlines do not want to take responsibility and leave the decision to the discretion of the pilot. In an attempt to ship samples collected in Colombia, LCS had to wait more than two months before receiving the samples. Only 25% of the samples could be successfully immortalized, most probably because of partial evaporation of liquid nitrogen in storage.

6. Conservation

Because many of the cell lines that will be established will be irreplaceable resources that will be studied for decades to come very special conservation procedures must be established. For example, it will be necessary periodically to monitor the viability of stored cell lines and necessary when they are frozen in a central repository to check for mycoplasma infection and, if present, treat them to eliminate it. It is also essential to establish a complete cell line lineage that will allow tracking of descendent cell cultures over a long period of time. This will provide a check against possible mix-ups or contamination of the cell lines that may happen in secondary repositories or parallel distribution schemes.

Conservation must also involve record keeping in a centralized data base accessible to all researchers of the characterization of the individual from whom each cell line has been established. Minimal criteria for such documentation have

to be established, but must involve name (to be kept confidential), stated ethnic group, birth place (including both name and geographic coordinates), birth place of grand-parents (desirable, but not essential), first language and dialect, known relationship to any other individuals in the collection, etc. This data base would necessarily indicate where the cell line and DNA samples were stored and how they would be available.

A third part of conservation would involve possible monitoring and "upgrading" of the samples. For example, the current sample of Chinese is not a geographically proscribed specific group, but rather somewhat more broadly representative of Han Chinese. While valuable for preliminary studies, it should be supplanted for most research studies by more specific collections, allowing studies of regional variation even among the Han. Similarly, the sample of Nasioi is small and contains some related individuals. If it is possible in the future to obtain samples from additional unrelated members of the same tribe, the related individuals should be relegated to a secondary status. While the collection should be upgraded over time, the original samples that are "eliminated" from the working collection should not be discarded if some research has used them.

7. An Integrated Database

Ideally, in a manner very analogous to the CEPH database but not necessarily using their system, it would be desirable for the data collected by many different laboratories on each individual to be collected in a centralized data base with public access. This would allow any researcher attempting to do any type of research - including linkage disequilibrium studies - to obtain the necessary raw genetic data. It would also serve to eliminate duplication of effort by making readily available an integrated summary of all that was already known about each sample. While enforcing submission of data to a common data base may be difficult, the concept needs discussion and the majority of laboratories generating large amounts of data will probably be quite willing to participate to their mutual advantage. The design and organization of such a database is a research project in itself that could be started utilizing the many thousands of individual typings already accumulated at Yale and Stanford.

8. Cell line definition

An essential component of the collection must be the "fingerprinting" of each cell line in the collection according to some standard criteria. While an occasional misidentified cell line represents little problem at the level of gene frequency studies, it can be a major problem at finer levels of molecular characterization. If a given researcher finds a Caucasian haplotype in a Melanesian sample, and concludes ancient gene flow or parallel "mutation", it is essential that the DNA (cell line) with the usual haplotype be confirmable as truly part of the Melanesian collection and not a Caucasian sample somehow mislabelled. Decisions on the "fingerprinting" procedure should be made early in the study and each cell line characterized as soon as possible after it is established. It is not necessary that uniqueness within the collection be established; a low frequency of each fingerprint would be sufficient (i.e., 1:1,000-1:10,000). The ideal characterization would be systems that can be tested by simple Southern blotting and by PCR-based techniques. The highly polymorphic systems currently in use in forensics may not be ideal for such characterization because of the precision required for measuring fragment sizes. However, whatever systems are used, they should also be defined on a reference cell line so that individual laboratories have a known sample on which to validate their methodologies. The standard cell line being used by the U.S. National Bureau of Standards in forensics is a logical choice.

9. Production of PCR libraries

Libraries can be prepared from small amounts of material. A computer catalogue, and/or a repository of already existing and new samples should be made available to research workers.